Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 25 (2016) 208 – 215

Procedia Technology

Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology (RAEREST 2016)

An effective multi-clustering anonymization approach using discrete component task for non binary high dimensional data spaces

Arun Shalin L.V*,Dr.K.Prasadh[b]

* Research Scholor,ManonmaniumSundarnar University,Tirunelveli,India

[b] *Principal,Mookambika Institute of Technology,Kerala,India*

Abstract

Clustering in common is a process of grouping elements together, so that the elements assigned to the same cluster are more comparable to each other than the remaining data points. Certain difficulties related to dealing with high dimensional data are ubiquitous and abundant. Research works conducted using anonymization method for high dimensional data spaces failed to address the problem related to dimensionality reduction for non binary databases. In this paper, Discrete Component Task Specific Multi-Clustering (DCTSM) approach is presented for dimensionality reduction on non binary database. To start with the analysis of attribute in the non binary database takes place and the process of projecting clusters identifies sparseness degree of dimensions. Then with the quantum distribution on multi cluster dimension, the solution for relevancy of attribute and redundancy on non-binary data spaces is provided. As a result, dimensionality reduction on non binary data leads to performance improvement on the basis of tag based feature. Multi clustering tag based feature reduction extracts individual features and are correspondingly replaced by the equivalent feature clusters (i.e.) tag clusters. During training, the DCTSM approach, multi clusters are used instead of the individual tag features and then during decoding the individual features are replaced by the corresponding multi clusters. To measure the effectiveness of the method, experiments are conducted on existing anonymization method for high dimensional data spaces and compared with the DCTSM approach using Statlog German Credit Data Set. DCTSM approach obtained results of 7.05 % improved accuracy and was observed that it took minimal time during tag feature extraction and resulted in lesser error rate.

Peer-review under responsibility of the organizing committee of RAEREST 2016

*Keywords:*High-Dimensional Data Space;Non-Binary Database;Discrete Component Task Specific;Quantum Distribution; Dimensionality Reduction

Peer-review under responsibility of the organizing committee of RAEREST 2016

doi:10.1016/j.protcy.2016.08.099

1. Introduction

In the recent years, different types of clustering algorithms are introduced which are approximately separated into four types ranging from projection, hierarchical, density-based to subspace algorithms. The different types of algorithm as mentioned investigate the clusters in lower-dimensional projection of the original data. It is normally favored when dealing with information that is high dimensional. Motivated by the fact of high dimension, with the preference of observation that has more dimensions regularly leads to the so called curse of dimensionality, where the performance of many normal machine learning algorithms becomes impaired. This is frequently due to two pervasive effects such as the empty space incident and deliberation of distances.

The term 'curse of dimensionality' refers to the fact that all high dimensional data sets tend to be sparse, because the number of points necessary to symbolize any distribution grows exponentially with the number of dimensions. This results in bad density estimates for high-dimensional data, causing complexity for density based approaches on non- binary database. The latter is a rather counterintuitive property of high dimensional data point representations, where all distances between data points tend to turn out to be harder to differentiate with the increase in dimensionality which is omnipresent and copious.

Novel anonymization methods for sparse high-dimensional data [1] were based on estimated Nearest Neighbor (NN) search in high-dimensional spaces, which was evaluated using Locality Sensitive Hashing (LSH). The data transformation involved in it extracts the establishment using the underlying reduction into a band matrix and gray encoding-based sorting. These band matrixes and gray encoding made the establishment of anonymization in groups resulting in lesser information loss with the help of an efficient linear-time heuristic but problem related to non binary databases was not solved. Anonymization methods for sparse high-dimensional data do not use dimensionality reduction techniques for more effectual anonymization.

The idea of selecting subset of good features with high variance and feature subset selection are proved to be certain efficient methods for dimensionality reduction. With the selection of good feature, the irrelevant data are removed that increases the accuracy related to learning and maximizing the comprehensibility for non-binary database. The feature subset selection methods for non binary database are divided into four types namely, extensive category explicitly embedded, wrapper, filter, and hybrid techniques. The embedded methods integrate feature selection as a part of the training process and are usually precise, and therefore proved to be more efficient than the other three types.

On the basis of the aforementioned techniques and methods applied, the proposed work uses Discrete Component Task Specific Multi-Clustering (DCTSM) approach for non binary high dimensional data spaces to improve accuracy. DCTSM first clusters different types of pertinent attributes to recognize the constituent records of it. The problem of projected clustering is addressed by identifying the clusters and its appropriate attributes extracted from statlog german credit dataset. Subsequently, discrete dimensional projection clusters using the quantum distribution model are evolved that also considers the problem related to attribute relativity and redundancy.

DCTSM approach offers a multi cluster formation based on objective function and evolve a discrete dimensional projection clusters. Finally, dimensionality of the data is reduced by ignoring the lower Eigen value components. On the other hand, DCTSM approach uses the class information to perform a projection of the features which best separate two or more classes.

Experiments using datasets Statlog German Credit Data Set extracted from UCI repository confirm that, the DCTSM approach facilitates higher level of accuracy and also the multi-clustering process is considerably efficient. Empirical studies show that the adoption of the DCTSM approach improves the level of accuracy and minimizing the error rate compared to the significantly more efficient state of art technique. The contribution of Discrete Component Task Specific Multi-Clustering (DCTSM) approach on non binary database for dimensionality reduction mining includes the following:

(1) To identify sparseness degree of dimensions using Discrete Component Task Specific Multi-Clustering (DCTSM) approach
(2) To provide solution provide solution for relevancy of attribute and redundancy on non-binary data spaces

(3) To reduce the dimensionality on the basis of tag based feature
(4) To apply multi clusters instead of the individual tag features and then during decoding the individual features are replaced by the corresponding multi clusters to reduce the error rate.

2. Materials and Method

In this section, we propose to adapt an approach for effective reduction of dimensionality based on tags in the non-binary database called as Discrete Component Task Specific Multi-Clustering. We describe the attribute relevance analysis, the model of quantum distribution and tag specific feature extraction for precise dimensional projection. Also we propose an algorithm for DCTSM approach. The architecture diagram of the DCTSM approach for reducing the dimensionality on non binary high dimensional data spaces is illustrated in figure 1.

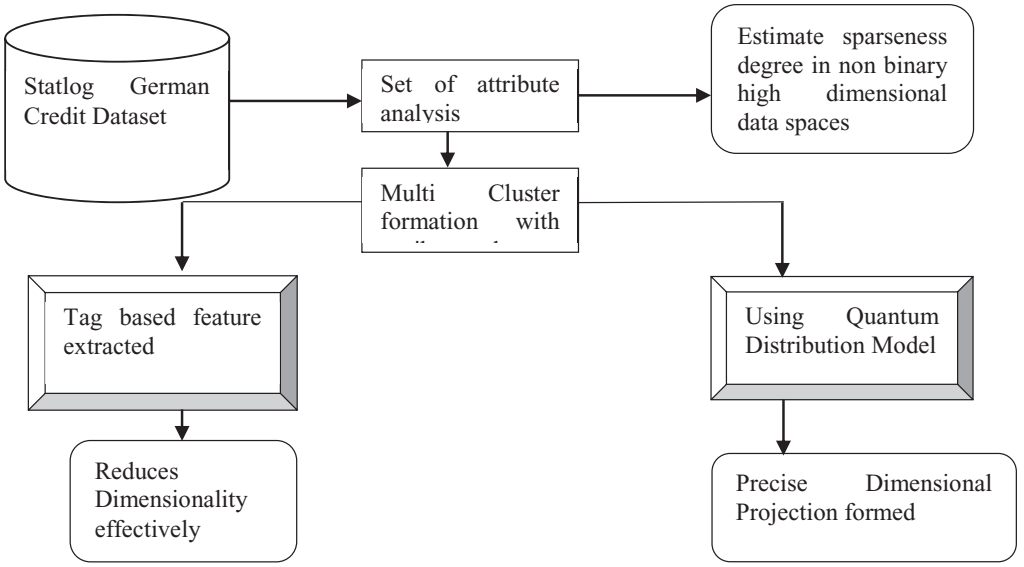

Figure 1 Diagrammatic representation of the DCTSM approach

Figure 1 illustrates the diagrammatical representation of the DCTSM approach which is divided into three phases. During the first and initial phase, analysis of attribute in the non-binary database takes place and the process of projecting clusters are deployed to identify sparseness degree even on the small subset of dimensions. Then, once the analysis of non-binary database is accomplished, the second phase develops a multi cluster dimension on quantum distribution for recognizing the relevancy of attribute and redundancy on non-binary data spaces. Finally, the third phase extracts the feature based on the tag to reduce the dimensionality on non-binary data using the attributes extracted from the UCI repository, Statlog German Credit Dataset

3.1 DCTSM Attribute Relevance Analysis

To facilitate attribute relevance analysis, we first recognize all proportions in a statlog german credit dataset which display certain level of cluster composition by determining dense regions and identify their position with the help of correct measurement. The fundamental hypothesis for attribute relevance analysis phase is that in scenarios including estimated clustering, a cluster contain appropriate proportions in which the projection of every point of the cluster is in an adequate number of further expected points, and this notion of ”proximity” is qualified with all the proportions. The dimensions that are further symbolized are then used as the probable aspirants for appropriate proportions of the clusters.

In DCTSM approach, consider DB a statlog german credit dataset of n-dimensional points, where the attributes is denoted by $A = \{A_1, A_2,\ldots,A_n\}$ and a set of N non-binary data points, where $xi = (x_{i1}, x_{i2},\ldots x_{in})$. Each x_{ij}

communicates with the assessment of data point x_i on attribute A_j where x_i is termed as one dimensional point. Each data point x_i fits in either one estimated cluster or to the position of outliers OUT.

For a given number of clusters, '*nc'*, acts as an input parameter, with an estimated cluster Cs, s = 1, 2, …, nc is termed as a pair (SP_s, SD_s) where SP_s is a division of non binary data points of database and SD_s is a division of dimensions of set of attributes A, such that the projections of the data points in SP_s beside every dimension in SD_s are directly grouped. The proportions in SP_s are termed as significant dimensions for the cluster C_s. The last proportions, (i.e.,) A-SD_s are termed as inappropriate dimensions for the cluster C_s. The cardinality of the position SD_s is indicated by d_s where d_s.d and n_s specifies the cardinality of the set SP_s where $n_s < N$.

With the help of attribute significance analysis, the sparseness level y_{ij} are determined for diverse proportions. The sparseness level y_{ij} are specified by

$$\text{yij} = \sum (\text{r} - \text{cij})\text{r} / \text{k}$$

where $r \in p_i^j(x_{ij})$. The least assessment of y_{ij} signifies solid region whereas the highest assessment signifies thin region of non-binary data. Likewise different y_{ij} values are determined for different spatial images for different dimensions which facilitate assessment of y_{ij} for every image, i.e., it simply perceive the dense regions. The images with better values of y_{ij} indicate thin regions whereas the attribute with less values of y_{ij} signify the opaque regions.

3.2 *Quantum Distribution Model in DCTSM approach*

Once the attribute relevance analysis is performed, the quantum distribution model is applied in DCTSM approach. Each cluster is precise in selecting the attribute value using DCTSM approach. Followed by this multiple clusters are formed for different types of attribute value and each attribute value of a cluster is a restricted mean chosen precisely from the domain. Each report in the multi cluster then follow the precise attribute values according to the data error rate. Quantum distribution is optimized with a given constraints and with variables that need to be minimized or maximized using programming techniques.

The DCTSM approach necessitates a quantum distribution model by simpler form so that an available computational objective function approach is used on non-binary database. This leads to a natural criterion for selecting the most excellent precise attribute value. Statlog german credit dataset consists of a set of incoming reports which are denoted by $Y1 \ldots Yi$.

Let us assume that the data point Yi is received during the time stamp Si followed by the assumption that the discrete dimensionality of the non binary database is 'h'. The 'h' dimensions of the report Yi are denoted by (y^1i. . . y^d i). In addition, each non-binary data point has an error associated with dissimilar dimensions. The error associated with the k^{th} dimension for non-binary data point Y_i is denoted by ψk (Yi).

Since different dimensions of data replicate diverse quantities, they correspond to different scales in DCTSM approach. In order to take the precise behavior of the different discrete dimensions into account, quantum distribution across different discrete dimensions is performed. DCTSM approach maintains the global statistics to compute global variances. These variances are used to scale the data over time with values.

In order to include the greater importance of recent data points in a developing stream, concept of an objective function f(s) quantifies the relative importance of the different data points over time. The objective function is drawn from the range (0, 1), and serves as a quantize factor for the relative importance of a given data point with a decreasing function that represents the objective of importance of a data point over time.

The objective function of quantum distribution model for DCTSM approach is the exponential objective function on non-binary database. The exponential objective function f(s) with parameter λ is defined as follows as a function of

$$f(s) = 2^{-\lambda \cdot s} \quad (1) \qquad \ldots\ldots\ldots\ldots\ldots \text{Eqn (2)}$$

The value of f(s) reduces by a factor of 2 every 1/λ time units and corresponds to the half of the function f(s).

3.4 Algorithm for DCTSM approach

The algorithm given below describes the steps performed in Discrete Component Task Specific Multi-Clustering (DCTSM):

Input: Let Aj denote the attribute values of Statlog German Credit Data Set 'S'. $Q_{min,}$ is minimum number of Quantum index of a selected attribute, 'MC' is multi cluster k^{th} dimensions of dataset, t occurs 'n' times in the corpus.

Output: Attribute analysis with sparseness degree and Dimensionality reduction on non-binary database.

Begin

// Analysis of attribute on 'S'

1: Compute the sparseness degree y_{ij};
2: Normalize y_{ij} in the interval [0, 1];
3: For m = 1 to m_max do
4: If (m = = 1) then
5: Estimate the parameters of the gamma distribution based on the probability
6: Compute the value of sparseness using Eqn 1
7: End If

// recognize relevancy of attribute and redundancy

8: For each cluster C, SelectAttrisVal(C, Q min)
9: BuildQDmodel($|B_{\min}|$, Q min)
10: While (Quantizeresult)
11: MC1 and MC2 are multiple clusters formed with objfunc
12: Various attribute value forms the new cluster Cn
13: Cn: = MC1, MC2……MCn
14: End While
15: Update Quantize result
16: End for each

// Dimensionality reduction on multi cluster

17: SelectAttrisVal(Cn, Q_{min})
18: If (Tag't')
19: Register_Prev vector representation on v_{i-1} or v_{i-2} tag set
20: Register _Next vector representation on v_{i+1} or v_{i+2} tag set
21: If (position 'pos')
22: Register_Prev vector representation on v_{k-1} or v_{k-2} tag set
23: Register _Next vector representation on v_{k+1} or v_{k+2} tag set
24: Tag vector v_k computed
25: Enf If
26: Calculate DC using Eqn (3) to reduce feature tags
27: End If

End

The above algorithmic step is used to identify the sparseness degree and dimensionality reduction on non-binary database using sparseness degree y_{ij} using Eqn (1) and the parameters are evaluated. In DCTSM approach, relevancy of attribute and redundancy are identified using the quantum distribution model. The quantum model produces result using the objective function after selecting the appropriate attribute value. The quantized result is updated. Then, dimensionality reduction in the non-binary database based on the tag based feature is obtained. Register_Prev and Register _Next are two form of vector representation for the positions v_{i-1} or $v_{i-2,}$ v_{i+1} or $v_{i+2,}$ v_{k-1} or $v_{k-2,}$ v_{k+1} or $v_{k+2.}$ Finally, v_k is computed and DC is obtained using the above equation to reduce the dimensionality.

3. Experimental Results

The main goal of the experiments presented here is to evaluate the capability of non-binary database to correctly identify dimensionality reduction in various situations. Experimental evaluation is conducted to estimate

the performance of the Discrete Component Task Specific Multi-Clustering Approach in high dimensional data spaces. The DCTSM approach is implemented in Java. The first phase achieves analysis of redundancy in the non-binary database using the Statlog German Credit Data Set. With the outcomes of the first phase, the objective of the second process is to provide solution for relevancy of attribute and redundancy, whereas the third phase efficiently reduces the dimensionality on multi cluster dimensions

Statlog German Credit Data Set classifies the publicly available data by a set of attributes as good or bad acclaim risks. Also Statlog German Credit Data Set contents come up with a cost matrix. In cost matrix, rows represent the actual classification and the columns denote the predicted classification. Statlog German Credit Data Set holds 20 attributes with 1000 instances. Only discrete dimension of attributes is precise to each multi cluster, and each multi cluster have different set of precise attribute value. An attribute is precise to a cluster if it helps to identify the member reports of it. This means the values at the precise attributes are distributed around some specific values in the cluster, while the reports of other clusters are less likely to have such values. Determining the multi clusters and their precise attribute value from a Statlog German Credit dataset is known as the discrete dimensional projected clusters.

4. Result and Discussion

Discrete Component Task Specific Multi-Clustering (DCTSM) approach is compared with the Anonymization methods for sparse high-dimensional data through locality-sensitive hashing (LSH) for measuring the accuracy, running time and error rate. Average Accuracy (AA) is needy on how data points (i.e.) information is collected, and is usually judged by comparing numerous capacity from the same or different multi clustered. The non-binary high dimensional database average accuracy is measured on dimensionality reduction

$$AA = (\textbf{Number of correctly classified tags})/(\textbf{Total number of tags on non binary data}).$$

The running time of tag feature extraction is defined as the amount of time taken to extract the tags from the non binary data points. It is measured in terms of milli seconds (ms). The Error Rate (ER) is the number of errors separated by the whole number of transferred data points on non binary database during specific time interval.It is measured as an estimated estimate of the error probability and precise for a long time interval and a high number of errors occurred on data space. ER is measured in terms of percentage (%).

The table (Table 1) given below shows the average accuracy measured with respect to the novel instances in the range of 100 to 700. Comparisons are made with the existing LSH technique to measure the effectiveness of the proposed DCTSM approach.

Novel instances	Average Accuracy (%)	
	LSH technique	DCTSM Approach
100	82	88
200	84	90
300	83	91
400	87	93
500	88	95
600	85	92
700	86	95

The table (Table 2) given below show the tag feature extraction running time with respect to the number of extracted tag features and elaborate comparisons are made with the existing LSH technique to measure the effect of running time using the proposed DCTSM approach.

No. of extracted tag features	Tag feature Extraction Running time (ms)	
	LSH technique	DCTSM Approach
50	787	771
80	767	752
120	778	762
150	884	866
190	895	881
250	873	861
320	921	903

Table 3 illustrates the error rate measured based on the non binary data dimensions and measured in terms of the percentage (%). The error rate in DCTSM approach is reduced to 10 – 20 % using the discrete component (DC). Each vector in DC is of dimension m + 1 corresponding to 'm' discrete component group (Mj). In addition with the introduction of indecisive DCTSM algorithm the possibility density function characterized the underlying behavior and decreases the error rate

Non binary data dimensions	Error rate (%)	
	LSH technique	DCTSM Approach
50	6.1	5.6
100	10.6	9.3
150	13.5	12.1
200	22.8	19.3
250	29.1	25.5
300	29.5	25.9
350	32.3	27.9

As a final point, dimensionality is reduced in the non binary high dimensional data space points. The ratio between the number of tag occurrences and its total number of occurrences in the quantity are used as a metric of DC task specific tag selection. Some tags occur only once in the training quantity and considered as a context tag. These tags with higher discrete component find place in the important tag list. But these tags which are not much frequent are removed for further reduction and improve the accuracy rate.

5. Conclusion

The performance of Discrete Component Task Specific Multi-Clustering approach initially analysis the attribute present in the non-binary statlog german credit dataset and the process of projecting clusters are used to identify sparseness degree of data dimensions. Then multi-cluster dimension provide relevancy of attribute and redundancy on non-binary data spaces using quantum distribution. As a final point, multi clustering tag based feature reduction takes individual features and are replaced by the equivalent feature clusters for dimensionality reduction. Recognition of an appropriate position based on tag feature in DCTSM approach is very imperative to give up the best performance for the dimensionality reduction. Generally the context and affix information are

useful in identification of the discrete component from a tag. The DCTSM approach gives up results with 7.05 % improved accuracy, minimal time taken for tag feature extraction, and lesser error rate. The effectiveness of the feature reduction based on tag approach is carried out in better way than the analogous attribute sets.

References

[1] Gabriel Ghinita, Panos Kalnis and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011

[2] Jung-Yi Jiang., Ren-Jia Liou., and Shie-Jue Lee., "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3,2011

[3] Bouguessa, M., Shengrui Wang., "Mining Projected Clusters in High-Dimensional Spaces," IEEE Transactions on Knowledge and Data Engineering, Volume: 21 , Issue: 4, 2009

[4] Deng Cai Chiyuan., Zhang Xiaofei He., "Unsupervised Feature Selection for Multi-Cluster Data," ACM Transactions on Knowledge Discovery from Data, 2010.

[5] Sharadh Ramaswamy., and Kenneth Rose., " Adaptive Cluster Distance Bounding for High-Dimensional Indexing," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[6] Mohamed Bouguessa., and Shengrui Wang., " Mining Projected Clusters in High-Dimensional Spaces," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 4, APRIL 2009

[7] Mohammad M. Masud., Jing Gao., Latifur Khan., Jiawei Han., Bhavani Thuraisingham., "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[8] Ahmed Abbasi., Stephen France., Zhu Zhang, and Hsinchun Chen., "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

[9] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang., and Ming-Syan Chen., "Density Conscious Subspace Clustering for High-Dimensional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 1, JANUARY 2010

[10] Qinbao Song., Jingjie N.i, and Guangtao Wang., "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

[11] Brian Quanz., Jun (Luke) Huan., and Meenakshi Mishra., "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 10, OCTOBER 2012

[12] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic., "The Role of Hubness in Clustering High-Dimensional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, REVISED JANUARY 2013

[13] Duc Thang Nguyen, Lihui Chen., and Chee Keong Chan., "Clustering with Multiviewpoint-Based Similarity Measure," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012

[14] Muhammad Aamir Cheema., Xuemin Lin, Wei Wang., Wenjie Zhang., and Jian Pei., "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 4, APRIL 2010

[15] Mehran Yazdi and Arash Golibagh Mahyari," A NEW 2-D FRACTAL DIMENSION ESTIMATION BASED ON CONTOURLET TRANSFORM FOR TEXTURE SEGMENTATION", The Arabian Journal for Science and Engineering, Volume 35, April 2010

[16] Z. M. Nopiah, M. H. Osman, S. Abdullah, M. N. Baharin," Application of a Multi-Objective Approach and Sequential Covering Algorithm to the Fatigue Segment Classification Problem", Arabian Journal for Science and Engineering, March 2014, Volume 39, Issue 3, pp 2165-2177